ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING SAS/PRINCOMP*

W. T. Federer, C. E. McCulloch and N. J. Miles-McDermott

BU-918-M November 1986

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal components analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal components analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT / PCP, and SYSTAT / FACTOR. We show here the results from SAS/PRINCOMP, Version 5.

---

* Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

1. INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations. For example, a data set may consist of n = 260 samples and m = 15 different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the m(m-1)/2 pairwise correlations and note which correlations are close to unity. When a group of variables are all highly inter-correlated, one may be selected for use and the others discarded or the sum of all the variables may be used. When the structure is more complex, the method of principal components analysis (PCA) becomes useful.

In order to use and interpret a principal components analysis there needs to be some practical meaning associated with the various principal components. In Section 2 we describe the basic features of principal components and in Section 3 we examine some constructed examples using SAS/PRINCOMP to illustrate the interpretations that are possible. In Section 4 we summarize our results.

2. BASIC FEATURES OF PRINCIPAL COMPONENT ANALYSIS

PCA can be performed on either the variances and covariances among the m variables or their correlations. One should always

check which is being used in a particular computer package program. SAS can use either the variances and covariances or the correlations but uses the correlations by default. First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $\underset{\sim}{X} = (X_1, X_2, \ldots, X_m)$, as follows:

$$PC_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m = \underset{\sim\sim}{X}b_1$$

$$PC_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m = \underset{\sim\sim}{X}b_2$$

$$\vdots$$

$$PC_m = b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mm}X_m = \underset{\sim\sim}{X}b_m \quad ,$$

where X_i have mean zero. In matrix notation,

$$\underset{\sim}{P} = (PC_1, PC_2, \ldots, PC_m) = \underset{\sim}{X}\,(\underset{\sim}{b}_1, \underset{\sim}{b}_2, \ldots, \underset{\sim}{b}_m) = \underset{\sim}{XB},$$

and conversely $\underset{\sim}{X} = \underset{\sim}{P}\,B^{-1}$.

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_i is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\ldots,m \quad .$$

Under this constraint, the b_{ij} in PC_1 are chosen so that PC_1 has maximal variance.

If we denote the variance of X_i by s_i^2 and if we define the total variance, T, as $T = \Sigma_{i=1}^{m} s_i^2$, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \ldots + var(PC_m) = T = \Sigma_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m

variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, $\mathbf{S_x}$. Let D be a diagonal matrix with entries being the eigenvalues, λ_i, of S_x arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = \text{var}(PC_i)$

(ii) $\text{trace}(\mathbf{S_x}) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} \text{var}(PC_i)$

(iii) $\text{corr}(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda}_i}{s_j}$

(iv) $\mathbf{S_x} = \mathbf{B'DB}$.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i}$$

PCA using the correlation martrix is different in these respects:

(i) The total "variance" is m, the number of variables. (It is not truly variance anymore.)

(ii) The correlation between PC_i and X_j is given by

$b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i}$. Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value.

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables would be used (PCA on the correlation matrix). To illustrate some of the above ideas, a number of examples have been constructed and these are described in Section 3. In each case, two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that in general, **PCA will not recover the original variables Z_1 and Z_2.** Both standardized and nonstandardized computations will be made.

3. EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following variance-covariance matrix (a variance-covariance matrix has the variance for the i^{th} variable in the i^{th} row and i^{th} column and the covariance between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column).

Variance-covariance matrix of Z_1 and Z_2

$$\begin{bmatrix} 11 & 0 \\ 0 & 85.8 \end{bmatrix}$$

Thus the variance of Z_1 is 11 and the covariance between Z_1 and Z_2 is zero. Also the total variance is 11 + 85.8 = 96.8.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. If PCA is performed on the variance-covariance

matrix then the SAS output is as follows (SAS control language for this example and all subsequent examples is in the appendix and the boldface print was typed on computer output to explain the calculation performed):

PCA1: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

11 OBSERVATIONS
2 VARIABLES

SIMPLE STATISTICS

	Z1	Z2
MEAN = $\bar{Z}_i$	0.000000	0.000000
ST DEV = s_i	3.316625	9.262829

COVARIANCES = s_{ij}

	Z1	Z2
Z1	11 = s_1^2	0 = s_{12}
Z2	0 = s_{21}	85.8 = s_2^2

TOTAL VARIANCE = 96.8 = T = 11 + 85.8

	EIGENVALUE ($\lambda_i = s_i^2$)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	85.80000	74.80000	0.88636	0.88636
PRIN2	11.00000	.	0.11364	1.00000

$\Sigma_{i=1}^{m} \lambda_i = T = 85.8 + 11.0 = 96.8$

proportion of variance explained by PC_i

EIGENVECTORS = $\underset{\sim}{b}_i$

PRIN1 = $\underset{\sim}{b}_1$ PRIN2 = $\underset{\sim}{b}_2$

Z1	b_{11} = 0.000000	b_{21} = 1.000000	$PC_i = b_{i1}Z_1 + b_{i2}Z_2$
Z2	b_{12} = 1.000000	b_{22} = 0.000000	$PC_1 = 0Z_1 + 1Z_2$

OBS	Z1	Z2	PRIN1 = PC_1	PRIN2 = PC_2
1	-5	15	15	-5 = 1(-5) + 0(15) = -5
2	-4	6	6	-4
3	-3	-1	-1	-3
4	-2	-6	-6	-2
5	-1	-9	-9	-1
6	0	-10	-10	0
7	1	-9	-9	1
8	2	-6	-6	2
9	3	-1	-1	3
10	4	6	6	4
11	5	15	15	5

We can interpret the results as follows:

1) The first principal component is

$$PC_1 = 0 \cdot X_1 + 1 \cdot X_2 = X_2$$

2) $PC_2 = 1 \cdot X_1 + 0 \cdot X_2 = X_1$

3) $Var(PC_1) = \text{eigenvalue} = 85.8 = Var(X_2)$

4) $Var(PC_2) = \text{eigenvalue} = 11.0 = Var(X_1)$

The PCs will be the same as the Xs whenever the Xs are uncorrelated. Since X_2 has the larger variance, it becomes the first principal component.

If PCA is performed on the correlation matrix we get slightly different results.

Correlation Matrix of Z_1 and Z_2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column. PCA in SAS would yield the following output:

PCA1: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

CORRELATIONS = r_{ij}

	X1	X2
X1	1.0000 = r_{11}	0.0000 = r_{12}
X2	0.0000 = r_{21}	1.0000 = r_{22}

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	1.000000	0.000000	0.500000	0.500000
PRIN2	1.000000	.	0.500000	1.000000

$\Sigma_{i=1}^{m} \lambda_i = m$

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1 = $\underset{\sim}{b}_1$	PRIN2 = $\underset{\sim}{b}_2$	
X1	b_{11} = 1.000000	b_{21} = 0.000000	$PC_i = b_{i1}Z_1/s_1 + b_{i2}Z_2/s_2$
X2	b_{12} = 0.000000	b_{22} = 1.000000	$PC_1 = 1Z_1/3.32 + 0Z_2/9.26$
			$= Z_1/3.32$

OBS	X1	X2	PRIN1	PRIN2	
1	-5	15	-1.5076	1.6194	=0(-5)/3.32+1(15)/9.26
2	-4	6	-1.2060	0.6478	
3	-3	-1	-0.9045	-0.1080	
4	-2	-6	-0.6030	-0.6478	
5	-1	-9	-0.3015	-0.9716	
6	0	-10	0.0000	-1.0796	
7	1	-9	0.3015	-0.9716	
8	2	-6	0.6030	-0.6478	
9	3	-1	0.9045	-0.1080	
10	4	6	1.2060	0.6478	
11	5	15	1.5076	1.6194	

The principal components are again the Xs (standardized Zs) themselves, but the eigenvalues (var(PCs)) are unity since the variables have been standardized first.

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. If the analysis is performed on the variance-covariance matrix using SAS the results are:

PCA2: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

11 OBSERVATIONS
3 VARIABLES

SIMPLE STATISTICS

	X1	X2	X3
MEAN	0.000000	0.000000	0.000000
ST DEV	3.316625	6.633250	9.262829

COVARIANCES

	X1	X2	X3
X1	11	22	0
X2	22	44	0
X3	0	0	85.8

TOTAL VARIANCE = 140.8

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{1+1}$)	PROPORTION	CUMULATIVE
PRIN1	85.80000	30.80000	0.60938	0.60938
PRIN2	55.00000	55.00000	0.39062	1.00000
PRIN3	0.00000	.	0.00000	1.00000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1	PRIN2	PRIN3
X1	0.000000	0.447214	0.894427
X2	0.000000	0.894427	-.447214
X3	1.000000	0.000000	0.000000

$$PC_3 = .894X_1 - .447X_2 + 0X_3$$

OBS	X1	X2	X3	PRIN1=PC_1	PRIN2=PC_2	PRIN3=PC3
1	-5	-10	15	15	-11.180	1.99840E-15
2	-4	-8	6	6	-8.944	1.55431E-15
3	-3	-6	-1	-1	-6.708	1.33227E-15
4	-2	-4	-6	-6	-4.472	8.88178E-16
5	-1	-2	-9	-9	-2.236	4.02456E-16
6	0	0	-10	-10	0.000	0
7	1	2	-9	-9	2.236	-4.02456E-16
8	2	4	-6	-6	4.472	-8.88178E-16
9	3	6	-1	-1	6.708	-1.33227E-15
10	4	8	6	6	8.944	-1.55431E-15
11	5	10	15	15	**11.180**	-1.99840E-15

$$11.180 = .447(5) + .894(10) + 0(15)$$

Analyzing the correlation matrix gives the following results:

PCA2: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

CORRELATIONS

	X1	X2	X3
X1	1.0000	1.0000	0.0000
X2	1.0000	1.0000	0.0000
X3	0.0000	0.0000	1.0000

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	2.000000	1.000000	0.666667	0.666667
PRIN2	1.000000	1.000000	0.333333	1.000000
PRIN3	0.000000	.	0.000000	1.000000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1	PRIN2	PRIN3
X1	0.707107	0.000000	-.707107
X2	0.707107	0.000000	0.707107
X3	0.000000	1.000000	0.000000

OBS	X1	X2	X3	PRIN1 =PC_1	PRIN2 =PC_2	PRIN3 =PC_3
1	-5	-10	15	-2.1320	1.6194	-4.44089E-16
2	-4	-8	6	-1.7056	0.6478	-3.74700E-16
3	-3	-6	-1	-1.2792	-0.1080	-2.77556E-16
4	-2	-4	-6	-0.8528	-0.6478	-1.80411E-16
5	-1	-2	-9	-0.4264	-0.9716	-8.32667E-17
6	0	0	-10	0.0000	-1.0796	0
7	1	2	-9	0.4264	-0.9716	8.32667E-17
8	2	4	-6	0.8528	-0.6478	1.80411E-16
9	3	6	-1	1.2792	-0.1080	2.77556E-16
10	4	8	6	1.7056	0.6478	3.74700E-16
11	5	10	15	2.1320	1.6194	4.44089E-16

$$2.1320 = .707107\left[\frac{X_1-\bar{X}_1}{s_1}\right] + .707107\left[\frac{X_2-\bar{X}_2}{s_2}\right] + 0\left[\frac{X_3-\bar{X}_3}{s_3}\right]$$

$$= .707107\left[\frac{5-0}{3.316625}\right] + .70107\left[\frac{10-0}{6.63325}\right] + 0\left[\frac{15-0}{9.262829}\right] .$$

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since given X_1 and X_3, X_2 is computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the sum of the variances, i.e.,

$$11 + 44 + 85.8 = 140.8$$

and

$$1 + 1 + 1 = 3 \ .$$

iv) For the variance-covariance analysis, the ratio of the coefficients of X_1 and X_2 in PC_2 is the same as the ratio of the variables themselves (since $X_2 = 2X_1$).

v) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

vi) The coefficients help to relate the variables and the PCs. In the variance-covariance analysis,

$$\text{Corr}(PC_2, X_1) = \frac{(\text{coefficient of } X_1 \text{ in } PC_2)\sqrt{\text{var}(PC_2)}}{\sqrt{\text{var}(X_1)}}$$

$$= \frac{b_{21}\sqrt{\lambda_2}}{s_1}$$

$$= \frac{.447214\sqrt{55}}{3.16625}$$

$$= 1 \ .$$

In the correlation analysis,

$$\text{Corr}(PC_1, X_1) = b_{11}\sqrt{\lambda_1}$$

$$= .707107\sqrt{2}$$

$$= 1 \ .$$

Thus, in both these cases, the variable is perfectly correlated with the PC.

vii) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, in the variance-covariance analysis, X_3 is clearly given by PC_1 . X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_2/\sqrt{5}$$

$$X_2 = 2 \cdot PC_2/\sqrt{5} \quad .$$

As a numerical example,

$$-5 = -11.180/\sqrt{5} \quad .$$

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1+5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. The analyses for the variance-covariance matrix (unstandardized analysis) and correlation matrix (standardized analysis) are given below:

PCA3: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

11 OBSERVATIONS
4 VARIABLES

SIMPLE STATISTICS

	X1	X2	X3	X4
MEAN	0.000000	10.00000	15.00000	0.000000
ST DEV	3.316625	6.63325	9.94987	9.262829

COVARIANCES

	X1	X2	X3	X4
X1	11	22	33	0
X2	22	44	66	0
X3	33	66	99	0
X4	0	0	0	85.8

TOTAL VARIANCE = 239.8

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	154.0000	68.2000	0.6422	0.6422
PRIN2	85.8000	85.8000	0.3578	1.0000
PRIN3	0.0000	0.0000	0.0000	1.0000
PRIN4	0.0000	.	0.0000	1.0000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1=PC_1	PRIN2=PC_2	PRIN3=PC_3	PRIN4=PC_4
X1	0.267261	0.000000	0.358569	0.894427
X2	0.534522	0.000000	0.717137	-.447214
X3	0.801784	0.000000	-.597614	0.000000
X4	0.000000	1.000000	0.000000	0.000000

OBS	X1	X2	X3	X4	PRIN1	PRIN2	PRIN3	PRIN4
1	-5	0	0	15	-18.708	15	2.44249E-15	-6.66134E-16
2	-4	2	3	6	-14.967	6	1.99840E-15	-4.44089E-16
3	-3	4	6	-1	-11.225	-1	1.77636E-15	-4.44089E-16
4	-2	6	9	-6	-7.483	-6	1.11022E-15	-2.22045E-16
5	-1	8	12	-9	-3.742	-9	6.66134E-16	-1.38778E-16
6	0	10	15	-10	0.000	-10	0	0
7	1	12	18	-9	3.742	-9	-6.66134E-16	1.38778E-16
8	2	14	21	-6	7.483	-6	-1.11022E-15	2.22045E-16
9	3	16	24	-1	11.225	-1	-1.77636E-15	4.44089E-16
10	4	18	27	6	14.967	6	-1.99840E-15	4.44089E-16
11	5	20	30	15	18.708	15	-2.44249E-15	6.66134E-16

$$18.708 = .267(X_1-\bar{X}_1) + .535(X_2-\bar{X}_2) + .802(X_3-\bar{X}_3) + 0(X_4-\bar{X}_4)$$

$$= .267(5) + .535(20-10) + .802(30-15)$$

PCA3: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

CORRELATIONS

	X1	X2	X3	X4
X1	1.0000	1.0000	1.0000	0.0000
X2	1.0000	1.0000	1.0000	0.0000
X3	1.0000	1.0000	1.0000	0.0000
X4	0.0000	0.0000	0.0000	1.0000

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	3.000000	2.000000	0.750000	0.750000
PRIN2	1.000000	1.000000	0.250000	1.000000
PRIN3	0.000000	0.000000	0.000000	1.000000
PRIN4	-.000000	.	-.000000	1.000000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.577350	0.000000	0.707107	0.408248
X2	0.577350	0.000000	-.707107	0.408248
X3	0.577350	0.000000	0.000000	-.816497
X4	0.000000	1.000000	0.000000	0.000000

OBS	X1	X2	X3	X4	PRIN1=PC_1	PRIN2=PC_2	PRIN3=PC_3	PRIN4=PC_4
1	-5	0	0	15	-2.6112	1.6194	-2.22045E-16	-2.22045E-16
2	-4	2	3	6	-2.0889	0.6478	-1.80411E-16	-5.55112E-17
3	-3	4	6	-1	-1.5667	-0.1080	-1.38778E-16	-5.55112E-17
4	-2	6	9	-6	-1.0445	-0.6478	-9.71445E-17	-2.77556E-17
5	-1	8	12	-9	-0.5222	-0.9716	-4.16334E-17	0
6	0	10	15	-10	0.0000	-1.0796	0	0
7	1	12	18	-9	0.5222	-0.9716	4.16334E-17	0
8	2	14	21	-6	1.0445	-0.6478	9.71445E-17	2.77556E-17
9	3	16	24	-1	1.5667	-0.1080	1.38778E-16	5.55112E-17
10	4	18	27	6	2.0889	0.6478	1.80411E-16	5.55112E-17
11	5	20	30	15	2.6112	1.6194	2.22045E-16	2.22045E-16

$$2.6112 = .577\left[\frac{X_1-\bar{X}_1}{s_1}\right] + .577\left[\frac{X_2-\bar{X}_2}{s_2}\right] + .577\left[\frac{X_3-\bar{X}_3}{s_3}\right] + 0\left[\frac{X_4-\bar{X}_4}{s_4}\right]$$

$$= .577\left[\frac{5-0}{3.317}\right] + .577\left[\frac{20-10}{6.633}\right] + .577\left[\frac{30-15}{9.950}\right] + 0\left[\frac{15-0}{9.263}\right]$$

For the variance-covariance analysis, the coefficients in PC_1 are in the same ratio as their relationship to Z_1. In the correlation analysis X_1, X_2 and X_3 have equal coefficients. In both analyses, as expected, the total variance is equal to the sum of the variances for the PCs. In both cases two PCs, PC_3 and PC_4, have zero variance; in the correlation analysis the PCs are identically zero but in the variance-covariance analysis they are constant, but not zero.

Example 4. In this example we take more complicated combinations of Z_1 and Z_2.

$$X_1 = Z_1$$

$$X_2 = 2Z_1$$

$$X_3 = 3Z_1$$

$$X_4 = Z_1/2 + Z_2$$

$$X_5 = Z_1/4 + Z_2$$

$$X_6 = Z_1/8 + Z_2$$

$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are colinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1. The PCAs for the variance-covariance and correlation matrices are given below:

PCA4: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

11 OBSERVATIONS
7 VARIABLES

SIMPLE STATISTICS

	X1	X2	X3	X4
MEAN	0.000000	0.000000	0.000000	0.000000
ST DEV	3.316625	6.633250	9.949874	9.410101

	X5	X6	X7
MEAN	0.000000	0.000000	0.000000
ST DEV	9.299866	9.272102	9.262829

COVARIANCES

	X1	X2	X3	X4
X1	11	22	33	5.5
X2	22	44	66	11
X3	33	66	99	16.5
X4	5.5	11	16.5	88.55
X5	2.75	5.5	8.25	87.175
X6	1.375	2.75	4.125	86.487
X7	0	0	0	85.8

	X5	X6	X7
X1	2.75	1.375	0
X2	5.5	2.75	0
X3	8.25	4.125	0
X4	87.175	86.487	85.8
X5	86.487	86.144	85.8
X6	86.144	85.972	85.8
X7	85.8	85.8	85.8

TOTAL VARIANCE = 500.8094

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	347.0151	193.2208	0.6929	0.6929
PRIN2	153.7943	153.7943	0.3071	1.0000
PRIN3	0.0000	0.0000	0.0000	1.0000
PRIN4	0.0000	0.0000	0.0000	1.0000
PRIN5	0.0000	0.0000	0.0000	1.0000
PRIN6	0.0000	0.0000	0.0000	1.0000
PRIN7	0.0000	.	0.0000	1.0000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.025018	0.264786	0.005297	0.081529
X2	0.050035	0.529573	0.077591	0.798545
X3	0.075053	0.794359	0.000856	-.563782
X4	0.504819	0.027439	-.306727	0.100525
X5	0.498565	-.038757	-.453794	-.106286
X6	0.495438	-.071855	0.830118	-.087701
X7	0.492310	-.104954	-.069597	0.093462

	PRIN5	PRIN6	PRIN7
X1	-.033880	0.445673	0.850185
X2	0.161253	-.046780	-.212517
X3	-.024221	-.187931	-.098000
X4	-.757301	0.198440	-.165332
X5	0.624127	0.348492	-.147391
X6	0.053236	0.202704	-.093099
X7	0.079938	-.749637	0.405822

OBS	X1	X2	X3	X4	X5	X6	X7
1	-5	-10	-15	12.5	13.75	14.375	15
2	-4	-8	-12	4.0	5.00	5.500	6
3	-3	-6	-9	-2.5	-1.75	-1.375	-1
4	-2	-4	-6	-7.0	-6.50	-6.250	-6
5	-1	-2	-3	-9.5	-9.25	-9.125	-9
6	0	0	0	-10.0	-10.00	-10.000	-10
7	1	2	3	-8.5	-8.75	-8.875	-9
8	2	4	6	-5.0	-5.50	-5.750	-6
9	3	6	9	0.5	-0.25	-0.625	-1
10	4	8	12	8.0	7.00	6.500	6
11	5	10	15	17.5	16.25	15.625	15

OBS	PRIN1=PC_1	PRIN2=PC_2	PRIN3=PC_3	PRIN4=PC_4
1	25.921	-21.332	2.22045E-16	-1.33227E-15
2	8.790	-15.937	-4.71845E-16	-1.44329E-15
3	-4.359	-10.918	-4.02456E-16	-1.34615E-15
4	-13.525	-6.275	-4.16334E-16	-9.71445E-16
5	-18.709	-2.009	-5.13478E-16	-5.41234E-16
6	-19.911	1.881	-2.49800E-16	-2.49800E-16
7	-17.131	5.395	-2.91434E-16	4.16334E-17
8	-10.368	8.533	2.77556E-17	6.38378E-16
9	0.377	11.294	4.02456E-16	1.29063E-15
10	15.104	13.679	6.38378E-16	1.77636E-15
11	33.813	15.688	1.33227E-15	2.44249E-15

OBS	PRIN5=PC_5	PRIN6=PC_6	PRIN7=PC_7
1	-2.44249E-15	5.55112E-15	-2.44249E-15
2	-1.73472E-15	3.55271E-15	-2.44249E-15
3	-1.08247E-15	2.49800E-15	-2.41474E-15
4	-5.41234E-16	1.55431E-15	-2.22045E-15
5	1.24900E-16	4.44089E-16	-1.77636E-15
6	6.93889E-16	-6.66134E-16	-6.66134E-16
7	7.35523E-16	-1.33227E-15	-2.22045E-16
8	1.12410E-15	-1.77636E-15	8.88178E-16
9	1.16573E-15	-2.63678E-15	2.24820E-15
10	1.16573E-15	-3.33067E-15	3.55271E-15
11	4.44089E-16	-3.33067E-15	5.10703E-15

$$0 = -.0339(5) + .161(10) - .0242(15) - .757(17.5) + .624(16.25) + .0532(15.625) + .0799(15)$$

PCA4: USING CORRELATION MATRIX (STANDARDIZED VARIABLES)
PRINCIPAL COMPONENT ANALYSIS

CORRELATIONS

	X1	X2	X3	X4
X1	1.0000	1.0000	1.0000	0.1762
X2	1.0000	1.0000	1.0000	0.1762
X3	1.0000	1.0000	1.0000	0.1762
X4	0.1762	0.1762	0.1762	1.0000
X5	0.0892	0.0892	0.0892	0.9961
X6	0.0447	0.0447	0.0447	0.9912
X7	0.0000	0.0000	0.0000	0.9843

	X5	X6	X7
X1	0.0892	0.0447	0.0000
X2	0.0892	0.0447	0.0000
X3	0.0892	0.0447	0.0000
X4	0.9961	0.9912	0.9843
X5	1.0000	0.9990	0.9960
X6	0.9990	1.0000	0.9990
X7	0.9960	0.9990	1.0000

	EIGENVALUE (λ_i)	DIFFERENCE ($\lambda_i - \lambda_{i+1}$)	PROPORTION	CUMULATIVE
PRIN1	4.052167	1.104335	0.578881	0.578881
PRIN2	2.947833	2.947833	0.421119	1.000000
PRIN3	0.000000	0.000000	0.000000	1.000000
PRIN4	0.000000	0.000000	0.000000	1.000000
PRIN5	-.000000	0.000000	-.000000	1.000000
PRIN6	-.000000	0.000000	-.000000	1.000000
PRIN7	-.000000	.	-.000000	1.000000

EIGENVECTORS = $\underset{\sim}{b}_i$

	PRIN1	PRIN2	PRIN3	PRIN4
X1	0.144294	0.557325	0.390963	-.117472
X2	0.144294	0.557325	0.390963	-.117472
X3	0.144294	0.557325	-.731868	0.338165
X4	0.493342	-.068313	-.320125	-.802613
X5	0.486325	-.118813	0.055387	0.318377
X6	0.481330	-.144088	0.031700	0.219949
X7	0.475353	-.169177	0.228280	0.253212

	PRIN5	PRIN6	PRIN7
X1	0.707107	0.023960	-.036628
X2	-.707107	0.023960	-.036628
X3	-.000000	-.029031	0.133187
X4	-.000000	-.071177	0.014612
X5	0.000000	-.485671	-.639580
X6	-.000000	0.826522	-.122632
X7	0.000000	-.271895	0.745160

OBS	X1	X2	X3	X4	X5	X6	X7
1	-5	-10	-15	12.5	13.75	14.375	15
2	-4	-8	-12	4.0	5.00	5.500	6
3	-3	-6	-9	-2.5	-1.75	-1.375	-1
4	-2	-4	-6	-7.0	-6.50	-6.250	-6
5	-1	-2	-3	-9.5	-9.25	-9.125	-9
6	0	0	0	-10.0	-10.00	-10.000	-10
7	1	2	3	-8.5	-8.75	-8.875	-9
8	2	4	6	-5.0	-5.50	-5.750	-6
9	3	6	9	0.5	-0.25	-0.625	-1
10	4	8	12	8.0	7.00	6.500	6
11	5	10	15	17.5	16.25	15.625	15

OBS	PRIN1=PC_1	PRIN2=PC_2	PRIN3=PC_3	PRIN4=PC_4
1	2.2378	-3.2844	1.38778E-16	2.91434E-16
2	0.5425	-2.3045	8.32667E-17	9.71445E-17
3	-0.7368	-1.4322	7.80626E-18	2.94903E-17
4	-1.6003	-0.6677	-1.38778E-17	0
5	-2.0479	-0.0108	-6.93889E-17	-1.24900E-16
6	-2.0795	0.5384	-5.55112E-17	-2.77556E-17
7	-1.6953	0.9799	-6.93889E-17	-5.55112E-17
8	-0.8951	1.3137	-5.55112E-17	-5.55112E-17
9	0.3209	1.5399	-2.08167E-17	-3.46945E-17
10	1.9529	1.6584	-1.38778E-17	-2.77556E-17
11	4.0007	1.6692	-5.55112E-17	2.22045E-16

OBS	PRIN5=PC_5	PRIN6=PC_6	PRIN7=PC_7
1	-1.89468E-16	6.66134E-16	-4.44089E-16
2	-4.33274E-17	4.85723E-16	-6.66134E-16
3	-3.59650E-17	3.48679E-16	-7.49401E-16
4	-1.47248E-17	2.22045E-16	-7.07767E-16
5	-7.36240E-18	1.11022E-16	-6.10623E-16
6	9.24446E-33	-9.71445E-17	-4.57967E-16
7	7.36240E-18	-1.24900E-16	-1.38778E-16
8	1.47248E-17	-2.49800E-16	1.94289E-16
9	3.59650E-17	-3.50414E-16	6.38378E-16
10	4.33274E-17	-4.71845E-16	1.15186E-15
11	1.89468E-16	-6.93889E-16	1.99840E-15

$$0 = .707\left[\frac{5-0}{3.317}\right] - .707\left[\frac{10-0}{6.633}\right] + 0\left[\frac{15-0}{9.950}\right] + 0\left[\frac{17.5-0}{9.410}\right]$$

$$+ 0\left[\frac{16.25-0}{9.300}\right] + 0\left[\frac{15.625-0}{9.272}\right] + 0\left[\frac{15-0}{9.263}\right]$$

We note several things:

i) In both analyses there are only two eigenvalues that are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2 and PC_3 where the standardizrd X_1, X_2 and X_3 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCAs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

4. SUMMARY

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, $\underset{\sim}{X}$, with maximal variance:

$$\underset{\sim}{P} = \underset{\sim}{X}\underset{\sim}{B} .$$

This relationship can be inverted to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed - see example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structures in the Xs.

APPENDIX

Example 1: Control Language

Control language is typed in upper case and comments are bolded. Refer to SAS User's Guide: Statistics, Version 5 Edition, 1985, for program documentation.

```
DATA ONE;
TITLE PCA1: USING CORRELATION MATRIX (STANDARDIZED VARIABLES);
INPUT X1 X2;     ⇒ input variables
CARDS;     ⇒ signals SAS that data follow
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
PROC PRINCOMP OUT=FOUR;     ⇒ requests PCA on correlation matrix
                                with PCs being output to new data set*
PROC PRINT DATA=FOUR;     ⇒ prints out data
```

* SAS will compute the PCA on the correlation matrix unless otherwise directed. To request PCA on a variance-covariance matrix use the following procedural call:

```
PROC PRINCOMP COV OUT=FOUR
```

Example 2: Control Language

```
DATA ONE;
TITLE PCA2: USING VARIANCE-COVARIANCE MATRIX (UNSTANDARDIZED VARIABLES);
INPUT Z1 Z2;
X1=Z1;      ⟹ creates X1,X2, and X3
X2=2*Z1;
X3=Z2;
DROP Z1 Z2;
CARDS;
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
PROC PRINCOMP COV OUT=FOUR;
VAR X1 X2 X3;          ⇒ tells SAS which variables to use to compute PCA
PROC PRINT DATA=FOUR;
```

Example 3: Control Language

```
DATA ONE;
TITLE PC3: USING CORRELATION MATRIX (STANDARDIZED VARIABLES);
INPUT Z1 Z2;
X1=Z1;
X2=2*(Z1+5);
X3=3*(Z1+5);
X4=Z2;
DROP Z1 Z2;
CARDS;
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
PROC PRINCOMP OUT=FOUR COV;
VAR X1 X2 X3 X4;
PROC PRINT DATA=FOUR;
```

Example 4: Control Language

```
DATA ONE;
TITLE PC4: USING CORRELATION MATRIX (STANDARDIZED VARIABLES);
INPUT Z1 Z2;
X1=Z1;
X2=2*Z1;
X3=3*Z1;
X4=(Z1/2)+Z2;
X5=(Z1/4)+Z2;
X6=(Z1/8)+Z2;
X7=Z2;
DROP Z1 Z2;
CARDS;
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
PROC PRINCOMP OUT=FOUR;
VAR X1 X2 X3 X4 X5 X6 X7;
PROC PRINT DATA=FOUR;
PROC MEANS DATA=FOUR; VAR PRIN1-PRIN7;
```